学校编码：10384　　分类号_____密级_____

学号：X2012231170　　UDC_____

厦门大学

工　程　硕　士　学　位　论　文

# 基于 GPS 的骑行社交混合推荐信息系统的设计与实现

# Design and Implementation of Hybrid Recommender Information System for Cycling Social Network Based on GPS

林　赟

指导教师姓名：　王 备 战　教 授

专 业 名 称：　软　件　工　程

论文提交日期：　2014 年 10 月

论文答辩日期：　2014 年 10 月

学位授予日期：　2014 年　　月

指　导　教　师：_____

答辩会委员主席：_____

2014 年 10 月

# 厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

# 厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

# 摘 要

随着移动互联网的迅速发展，基于位置的社交网络正获得越来越多的关注，同时带来了新的研究课题。运动社交是其中重要的细分领域。与智能硬件的结合，赋予这个领域更大的发展空间。

目前在地理位置数据分析和推荐系统方面已经有不少研究成果。但针对运动社交方面的研究数目还很有限。本文通过对骑行运动用户的调研，提出用运动地理兴趣和运动水平对用户的骑行运动特征进行综合衡量的思路。主要内容包括：

1. 对DBSCAN算法在地理位置聚类计算方面提出了面向运动特征分析的改进思路。通过在地理位置的 GPS 坐标维度之外增加了热度维，可以分析得出区域内大众运动热点区域分布。以此为基础，可以分析得出运动轨迹和用户行为的运动地理兴趣特征。该算法可以在利用用户-位置矩阵进行协同过滤时起到显著的降维效果，有利于提高计算的效率，提高特定需求的推荐语义准确性。

2. 骑行社交用户向量空间模型设计。利用混合推荐的思路，提出了一个反应骑行社交特性的用户模型，利用向量空间模型对用户的运动特征进行分析，并提出用户的运动特征相似性计算方式。模型中的地理兴趣维度来自协同过滤方法经过聚类转换后的结果。由于运动地理兴趣相关的大量语义信息（路面情况、景物等）无法在地理位置本身中得到体现，而协同过滤方法无需依赖分析对象的先天知识，可以从交互行为本身的分析获得较好的结果。模型中的运动水平维度应用了基于内容的推荐方法，通过对轨迹分析处理后提取所需的运动指标特征。两者的有机结合，为骑行社交用户推荐建立了可扩展的理论框架。

3. 骑行社交潜在队友推荐系统原型。基于骑行社交用户模型设计，本文进一步完善了算法设计，并实现了一个骑行社交推荐系统的原型系统。系统利用真实的骑行运动数据进行实验运行，初步验证了系统的有效性。

**关键词**：运动社交；位置聚类；推荐系统

# Abstract

With the rapid development of mobile Internet, location-based social networks are gaining more and more attention, and inspiring new research topics as well. Sports Social Network is one of the important segments. Accompanied with the intelligent hardware explosion, this field possesses broad prospects.

Compared with the quantity of research on geographic data analysis and recommender system, the studies on sports social network are still very limited. Based on the survey of the cyclists, this dissertation proposes a comprehensive measure framework for the cyclist's sports features by the sports geographical interest and exercise level. Focused on CSN (Cycling Social Network), the main contents include:

1. Improvement for DBSCAN on geographic clustering computation to enhance sports feature analysis. In addition to the GPS coordinates, the 'density of points' dimension is employed to analyze the distribution of collective sports hotspots in the target area. On this basis, the geographic characteristics of sports interest can be applied to the feature of user behavior and the trajectories. The algorithm can play a significant dimension reduction effect on the User-Location matrix calculation in collaborative filtering, and helps to improve the efficiency of computing and the accuracy of semantic recommendation as well.

2. A CSN user modeling based on Vector Space Model, which consists of the geographical interest dimension and the exercise level dimension as well. Applying hybrid recommender system and vector space model theory, we design a CSN user model to measure the user similarity of sports features. Applying collaborative filtering techniques, the geographical interest dimension exploits the results of clustering transformation of users' historical locations. Since the semantic information of sports geographical interests (for example, the scenery along the road, etc.) could not be retrieved directly from the trajectory analysis, the collaborative filtering is the appropriate techniques to choose for it doesn't need to know anything about the items. Applying the content-based recommender techniques, the exercise level dimension

exploits the sports features by preprocessing the user's trajectories. The integration of two dimensions establishes an extendable theoretical framework for the CSN friend recommender system.

3. A CSN friend recommender prototype system. Based on the CSN user model, we implement the algorithm and a prototype of CSN friend recommender system. By evaluating with the real-world cycling GPS datasets, the effectiveness of the prototype system has undertook the preliminary verification.

**Key Words:** Sports Social Network; Location Clustering; Recommender System

# 目 录

# Contents

# 第一章 绪 论

## 1.1 研究背景

随着全社会健康意识的提升，现代人对运动健康的需求越来越高，以骑行为代表的各类运动领域正处在快速发展中。调查报告指出，健康和社交是骑行运动中最主要的需求之一。但在相关领域，用户对运动的迫切需求和自身专业知识的相对薄弱、运动社群的不完善、基础设施的不健全存在一定的落差，自发性运动往往流于盲目和无计划。如果能将个体性运动与群体性运动相结合，可以大大提升运动体验，改善运动效果。好的运动社区能提供良好的骑行社交体验，群体化运动在满足用户的社交需求的同时，也能增加用户的运动乐趣，并进而改善锻炼效果，从而促进该项运动在社会范围内的发展和成熟。

经过一段时间的快速发展，基于地理位置的社交网络正在进入一个领域细分的发展阶段。随着泛社交应用的空间逐渐被巨头占据，SNS 的发展机遇更多的出现在各类垂直应用领域。而随着移动互联网和智能硬件的迅猛发展，运动健康作为一个重要的应用领域，正在迎来一个快速发展的阶段。由艾媒咨询（iiMedia Research）发布的《2012-2013 中国可穿戴设备市场研究报告》显示，预计到 2015 年中国市场可穿戴设备市场出货量将超过 4000 万部，市场规模达到 114.9 亿元[1]，而运动健康就是其中倍受关注的市场领域之一。从高科技企业（如谷歌、三星、苹果等）到传统运动用品公司（如 Nike、Adidas、Garmin 等），各大厂商纷纷推出以手环、手表、自行车智能硬件等在内的各项智能硬件产品。随着各种传感器技术和穿戴式硬件的快速发展，所积累的数据量也在急速攀升。这些数据当中蕴含着重要的信息，从人们的行为模式、生活方式，到用户的生理信息、运动状况，都可能分析产生巨大的价值。运动社交作为其中的一个重要应用分支，不但具有广阔的发展空间，也必将成为海量数据挖掘和利用的重要切入点。

## 1.2 研究意义

随着生活水平和消费水平的提高，人们健康意识的增强，运动用品市场正在经历快速发展的阶段。以运动自行车为例，根据视点行业调研发布的《2012-2015

中国运动型自行车市场专项调查报告》，2012 年国内运动自行车市场需求已超过 175 亿。但就对用户的深入了解和数据挖掘而言，整个行业仍然处于相当传统的状态。厂商、经销商无从了解用户对产品的实际使用情况，与用户之间存在脱节现象。在整个制造业开始向服务业转型的大趋势中，行业资源难以实现有效的优化配置。随着移动互联网的快速发展，民用 GPS 设备的普及，记录个人运动经历已经成为运动用户的新习惯。以用户运动数据为核心的应用服务，正在逐渐形成一个新的产业服务和价值链条。骑行社交就是这个数据服务的具体方向之一。数据服务的基础是对用户的广义运动档案的建模。针对社交应用方向构建的运动用户模型，将成为这个用户运动档案的重要组成。本文将以用户运动相似模型的设计为基础，以骑行社交用户推荐为具体应用，结合轨迹数据分析技术和推荐系统的相关理论，基于真实的运动地理轨迹数据，构建一个骑行社交的应用系统原型。

基于运动地理的大数据分析，可以为某个用户找出最相似的用户作为潜在运动伙伴，让那些可能在现实生活中有相似的兴趣，多次经过相同的区域，却一直没有机会认识的人一起运动。骑行社交还有很多扩展应用的可能。例如，从潜在运动伙伴的历史轨迹中，可以查找出该用户没有去过的地点，并利用协同过滤的方法为他推荐新的运动区域，进一步扩展他的运动体验。骑行社交应用不但可以提升用户的运动体验，还可以鼓励更多的用户积极的参与到运动中来，积累更多的数据信息，形成良性循环，让运动数据分析产生更大的价值。

## 1.3 研究现状

在推荐系统、地理位置社交网络信息挖掘等方面，国内国际上都已经有不少针对性的研究成果。作为一个新的应用领域，目前国内外直接针对骑行社交领域的研究成果还不多。但是，在了解现有关联性理论研究的过程中，可以得到非常有价值的参考和启发。

由于 LBSNS 的研究需要实际数据样本支持，以 Foursquare 为代表的签到类数据为不少用户-用户推荐的应用研究提供了数据基础。文献[2]通过统计分析，发现不同用户访问地点的相似性，好友之间较高，陌生人之间较低，说明地理位置对访客具有某种筛选效应。文献以地理位置为基础设计相应算法，选择用户访

问过的位置为中间节点，将访问过该位置的非好友用户作为候选，再进行选择推荐，并验证了推荐效果的改善。

微软亚洲研究院的 GeoLife 项目在地理位置社交方面取得的进展得到了很大的关注[3]。GeoLife 项目实际跟踪了 178 位用户，历经四年时间收集大量数据，包含 18465 条轨迹，总长度超过 10 万米。 该数据集记录了用户种类繁多的户外活动，包括购物、远足、聚餐等活动。在得到用户的行为轨迹数据基础上，文献[4]以停留点（Stay Point）为核心概念，对用户访问位置聚类，找出停留点，进而利用时间关系找出用户的访问序列模式，以此获得用户的偏好特征。停留点是一个很直观的概念。用户往往在目的地位置停留较长时间，所以停留点是用户轨迹中具有关键语义特征的区域。使用基于密度的聚类方法，可以从轨迹中分析出停留点。该算法将用户的轨迹表示成为一系列 stay point 的集合，并使用 HGSM 的数据结构，将用户的历史数据构建为特有的层次有向图。每个层次的节点代表了用户访问过的区域，层次越高，粒度越大；边代表了区域访问的先后顺序。这个结构具有很强的表达能力，不但可以表达用户的历史位置，也可以反映用户在地理空间的先后顺序和活动区域的层次性。基于这个结构，比较用户的区域访问序列，就可以获取用户之间的关联性。如果用户在较精细的层次有较多的匹配，则他们的相似度较高。

近年来，一些基于语义的用户相似性计算方法逐渐引起关注。在文献[5]中，Alvares 等人提出一种将用户的历史位置信息进行语义化的方法，将轨迹和其所处的具体地理环境相结合，从而挖掘出用户的语义轨迹模式。Alvares 等人提出了 stop 这个概念用以表示轨迹中的关键点，进而使用 stop 的序列表示一条轨迹。通过将 stop 和地标位置进行对比，可以实现 stop 语义赋值，进而将轨迹进行语义化。他们提出 SMoT（Stops and Moves of Trajectories）算法，通过模式序列挖掘算法计算用户的频繁模式序列，并以此作为分析用户相似性的依据。然而由于 Stops 局限于预先给定的区域，如酒店、景区等，普遍情况下可获得的 stops 较少，表达语义的能力也因此受到很大的限制。

## 1.4 研究内容

骑行社交是 LBSNS 的一个重要应用领域。本文以骑行运动为切入，在深入

发掘用户需求的基础上，广泛了解国内外的研究现状，进而以推荐系统理论为基础，应用地理位置轨迹分析的相关技术，设计了一个适用于骑行社交的用户模型空间，构造相应的应用系统原型，并基于真实的骑行运动数据进行了运动用户社交推荐的分析尝试。

目前针对骑行社交 LBS 服务的研究数量还相当有限。本文以真实运动应用和运动数据为基础，对运动地理位置服务和一般 LBS 的差异特征进行了比较讨论，并提出一种可以反映运动应用特征的地理位置聚类算法，为用户模型和推荐系统的构建奠定了基础。同时，现有基于 LBS 的用户相似性计算研究，多从地理位置或行为语义两方面单独入手。本文在构建用户模型空间时，将两者进行了融合，基于混合推荐技术提出了针对性的设计。最后，本文搭建了一个具有潜在运动伙伴推荐功能的系统原型，并通过真实的运动数据验证了该设计的有效性。

## 1.5 论文结构

本文共七章，其组织结构如下：

第一章，绪论。主要介绍了选题背景、当前国内外相关研究应用现状，对本文选题的原因和意义进行了阐述，对论文的内容和结构安排进行了概述。

第二章，推荐系统相关理论。以推荐系统的用户模型为侧重点，对推荐过滤技术相关的理论进行了比较分析，并针对骑行社交用户模型构建的特殊性进行了讨论，进一步明确以混合推荐为主要思路进行用户模型设计的方向。

第三章，地理位置数据的信息提取。首先提出了用两种维度衡量用户骑行社交需求的设计，包括运动地理兴趣维度和运动水平维度。二者的信息都可以从地理位置数据的分析中获得。其次比较了运动地理数据服务和其他 LBS 服务在目标关注点和数据特性方面的差异。进而介绍和骑行社交相关的地理位置信息分析所涉及的理论和技术，并提出了具有骑行社交特性的地理位置数据聚类算法思路。

第四章，用户模型设计。综合 LBS 轨迹分析技术和推荐系统相关理论，提出了适用于骑行社交场景的用户模型框架设计，包括用户向量空间的维度设定、归一化、相似度计算等相关设计思路。

第五章，算法设计。针对本文提出的用户模型设计，提出了一种具体的算法实现方式，提供了包括墨卡托映射进行计算简化、应用 DBSCAN 进行运动热点